



DICKINSON LAW REVIEW
PUBLISHED SINCE 1897


Volume 125 | Issue 3

Spring 2021

The Absence or Misuse of Statistics in Forensic Science as a Contributor to Wrongful Convictions: From Pattern Matching to Medical Opinions About Child Abuse

Keith A. Findley

Follow this and additional works at: <https://ideas.dickinsonlaw.psu.edu/dlr>

 Part of the [Criminal Law Commons](#), [Criminal Procedure Commons](#), [Criminology and Criminal Justice Commons](#), [Evidence Commons](#), [Family Law Commons](#), [Forensic Science and Technology Commons](#), [Legal Writing and Research Commons](#), [Litigation Commons](#), [Medical Sciences Commons](#), [Science and Technology Law Commons](#), and the [Statistics and Probability Commons](#)

Recommended Citation

Keith A. Findley, *The Absence or Misuse of Statistics in Forensic Science as a Contributor to Wrongful Convictions: From Pattern Matching to Medical Opinions About Child Abuse*, 125 DICK. L. REV. 615 (2021). Available at: <https://ideas.dickinsonlaw.psu.edu/dlr/vol125/iss3/2>

This Article is brought to you for free and open access by the Law Reviews at Dickinson Law IDEAS. It has been accepted for inclusion in Dickinson Law Review by an authorized editor of Dickinson Law IDEAS. For more information, please contact lja10@psu.edu.

The Absence or Misuse of Statistics in Forensic Science as a Contributor to Wrongful Convictions: From Pattern Matching to Medical Opinions About Child Abuse

Keith A. Findley*

ABSTRACT

The new scrutiny that has been applied to the forensic sciences since the emergence of DNA profiling as the gold standard three decades ago has identified numerous concerns about the absence of a solid scientific footing for most disciplines. This article examines one of the lesser-considered problems that afflicts virtually all of the pattern-matching (or “individualization”) disciplines (largely apart from DNA), and even undermines the validity of other forensic disciplines like forensic pathology and medical determinations about child abuse, particularly Shaken Baby Syndrome/Abusive Head Trauma (SBS/AHT). That prob-

* Professor of Law, University of Wisconsin Law School, co-founder and President of the Center for Integrity in Forensic Sciences, co-founder of the Wisconsin Innocence Project, and past president of the Innocence Network. Many thanks to Alicia Carriquiry, Simon Cole, Jay Koehler, Dean Strang, Leila Schneps, and Bill Thompson for invaluable comments on earlier drafts of this essay.

lem is the absence or misuse of statistics. This article begins by applying basic statistical principles to pattern-matching disciplines to demonstrate how those disciplines have historically hidden or failed to reckon with the probabilistic nature of their judgments, and how, when they have acknowledged the probabilistic nature of their claims, they have often botched the statistical analyses. The article then does a deeper dive into showing how those same deficiencies apply to medical opinions about child abuse, particularly SBS/AHT.

TABLE OF CONTENTS

INTRODUCTION	616
I. MAKING STATISTICAL OR PROBABILISTIC CLAIMS WITHOUT ACKNOWLEDGMENT, OR WITHOUT DATA TO SUPPORT THOSE CLAIMS	618
A. <i>Hidden Probabilistic Claims</i>	618
B. <i>Revelations from DNA Evidence</i>	623
C. <i>The Microscopic Hair Analysis Example</i>	625
D. <i>The Shaken Baby Syndrome/Abusive Head Trauma Example</i>	628
II. MAKING OVERTLY STATISTICAL CLAIMS, BUT BOTCHING THE NUMBERS: MORE ON THE SBS/AHT EXAMPLE	630
A. <i>Miscalculating the Probabilities</i>	630
1. <i>Overlooking the Base Rate</i>	631
2. <i>Misunderstanding the Diagnosticity of the Medical Evidence</i>	632
3. <i>Misunderstanding How to Combine Base Rates and Relevance Ratios</i>	633
B. <i>Miscalculating the Rarity of Alternative Causes</i>	638
III. MAKING PROBABILISTIC CLAIMS THAT EXCEED THE AUTHORITY OF THE EXPERT AND INTRUDE ON THE PROVINCE OF THE JUDGE OR JURY	646
CONCLUSION	651

INTRODUCTION

One of the great ironies of the innocence movement is that, while forensic science launched the movement by exposing errors that led to wrongful convictions in serious cases, those same wrongful convictions in turn exposed the deeply and systemically flawed nature of forensic science.¹ The science that launched the inno-

1. See PRESIDENT'S COUNCIL OF ADVISORS ON SCIENCE AND TECHNOLOGY, EXEC. OFF. OF THE PRESIDENT, REPORT TO THE PRESIDENT: FORENSIC SCI-

cence movement was of course DNA profiling, which was first developed in the mid-1980s and led to the first DNA exoneration in the United States in 1989 and a growing number since.² But the study of the wrongful conviction cases, as well as the model of real scientific inquiry set by DNA, have together demonstrated the unscientific and sometimes unreliable nature of many, if not most, of the other forensic disciplines.³

Numerous case examples and a growing body of scientific and legal scholarship, as well as rigorous reviews of the forensic sciences by esteemed and authoritative bodies such as the National Academy of Sciences (NAS)⁴ and the President's Council of Advisors on Science and Technology (PCAST),⁵ have identified many of the weaknesses in the research base and methodologies underlying everything from bitemark evidence to fingerprints. The critiques of the forensic sciences are broad and extensive.

It is perhaps not surprising then, although still disconcerting, that the record produced by post-conviction innocence advocacy over the past few decades has shown that flawed forensic science evidence has been a leading contributor to the problem of wrongful convictions. Misapplied forensic science evidence played a part in 52 percent of the false convictions in which DNA evidence alone later proved innocence.⁶ In cases where any type of new evi-

ENCE IN CRIMINAL COURTS: ENSURING SCIENTIFIC VALIDITY OF FEATURE-COMPARISON METHODS 2 (2016) [hereinafter PCAST REPORT].

2. See Keith A. Findley, *Innocence Found: The New Revolution in American Criminal Justice*, in *CONTROVERSIES IN INNOCENCE CASES IN AMERICA* 3, 4 (Sarah Lucy Cooper ed., 1st ed. 2014); see also Keith A. Findley, *Defining Innocence*, 74 ALB. L. REV. 1157, 1188 (2011) (discussing the impact of the DNA cases on the innocence movement).

3. See Michael J. Saks & Jonathan J. Koehler, *The Coming Paradigm Shift in Forensic Identification Science*, 309 SCIENCE 892, 893 (2005) (“What was unexpected [from the study of the initial DNA-exoneration cases] is that erroneous forensic science expert testimony is the second most common contributing factor to wrongful convictions, found in 63% of those cases.”); Vanessa Meterko, *Strengths and Limitations of Forensic Science: What DNA Exonerations Have Taught Us and Where To Go From Here*, 119 W. VA. L. REV. 639, 639 (2016).

4. NAT'L RES. COUNCIL OF THE NAT'L ACADEMIES, *STRENGTHENING FORENSIC SCIENCE IN THE UNITED STATES: A PATH FORWARD* 183 (2009) [hereinafter NAS REPORT].

5. PCAST REPORT, *supra* note 1, at 25.

6. *Overturing Wrongful Convictions Involving Misapplied Forensics*, THE INNOCENCE PROJECT, <https://bit.ly/3rJgLRS> [<https://perma.cc/CL8S-XMBS>] (last visited Mar. 14, 2021). According to the Innocence Project, which maintains the data on DNA exonerations, “the misapplication of forensic science is defined as an instance in which forensic evidence (i.e., analysis and/or testimony) was used to associate, identify, or implicate someone who was later conclusively proven innocent with post-conviction DNA testing, thereby demonstrating that the original forensic evidence was incorrect.” Meterko, *supra* note 3, at 640.

dence—not just DNA—produced an exoneration, false or misleading forensic science played a part in 24 percent of the wrongful convictions, according to the National Registry of Exonerations' database of more than 2,700 exonerations since 1989.⁷

In this essay, I address one of the less-considered weaknesses that has been exposed by scrutiny of the forensic sciences, a weakness that extends beyond the pattern-matching disciplines to a wide range of other disciplines, and indeed implicates the way that the law incorporates expert evidence as a whole under Federal Rule of Evidence 702. That weakness is the profound misunderstanding and misuse of statistics and probabilistic claims, or even the failure to explicitly acknowledge and analyze what are fundamentally statistical and probabilistic claims. I set forth here a few examples of the non-application or mis-application of essential statistical analyses to illustrate the problem. In doing so, I focus on three particular problems: making implicit or explicit probabilistic claims without acknowledging the nature of the claims and without supporting data; making fundamental errors in the application of statistics to data; and making statistical claims in ways that improperly intrude on the province of our system's legal fact-finders—the jury and judge.

I. MAKING STATISTICAL OR PROBABILISTIC CLAIMS WITHOUT ACKNOWLEDGMENT, OR WITHOUT DATA TO SUPPORT THOSE CLAIMS

A. *Hidden Probabilistic Claims*

All of the pattern-matching or individualization⁸ forensic disciplines are at bottom disciplines that can only make probabilistic claims—that is, statistical claims about the probability that trace evidence left at a crime scene might have been left by a particular suspect. For years, however, forensic analysts have testified in ways that suggest not probabilities, but certainties.⁹

7. % *Exonerations by Contributing Factor*, NAT'L REGISTRY OF EXONERATIONS, <https://bit.ly/3lbVTju> [<https://perma.cc/B9GK-ZUPU>] (last visited Mar. 14, 2021).

8. "Individualization" as used in the forensic sciences denotes those disciplines that attempt to match patterns in evidence found at the crime scene to evidence associated with the defendant in order to provide evidence of identity, that is, to connect the crime scene evidence to an individual.

9. See Simon A. Cole, *Forensics Without Uniqueness, Conclusions Without Individualization: The New Epistemology of Forensic Identification*, 8 L., PROBABILITY & RISK 233, 236 (2009); Simon A. Cole, *Individualization Is Dead, Long Live Individualization! Reforms of Reporting Practices for Fingerprint Analysis in the United States*, 13 L., PROBABILITY & RISK 117, 128 (2014).

Famously, fingerprint analysts long operated under standards promulgated by their own trade groups that declared it *unethical* to testify to anything but certainties.¹⁰ Fingerprint analysts were instructed that they should either definitively include—even to the exclusion of all other persons on earth—or exclude a suspect (or simply say the prints were not interpretable, or inconclusive). In the late 1990s, for example, the FBI-sponsored Technical Working Group on Friction Ridge Science, Analysis, and Technology (TWGFAST), which was later renamed the Scientific Working Group on Friction Ridge Analysis, Study, and Technology (SWGFAST), repeatedly promulgated guidelines permitting just three acceptable conclusions from latent print comparison: individualization (or identification), exclusion, or inconclusive.¹¹ Delegates at the annual conference of the International Association for Identification (IAI) in 1979 voted to “state unanimously that friction ridge [fingerprint] identifications are positive, and officially oppose any testimony or reporting of possible, probable or likely friction ridge identification.”¹² The delegates went on to declare that any fingerprint analyst who reported findings in terms of “possible, probable or likely friction ridge identification shall be deemed to be engaged in conduct unbecoming such member,” and would be subject to charges of misconduct by the Association, possibly resulting in decertification.¹³ In other words, fingerprint examiners themselves decided that it was professional misconduct to admit to a probabilistic conclusion.

The practice of declaring an “identification” or a “match” without qualification was common, and remains the norm today, in many of the other pattern-matching disciplines as well. As Simon

10. See Jennifer L. Mnookin, *The Validity of Latent Fingerprint Identification: Confessions of a Fingerprinting Moderate*, 7 L., PROBABILITY & RISK 127, 139 (2008) (“At present, fingerprint examiners typically testify in the language of absolute certainty. Both the conceptual foundations and the professional norms of latent fingerprinting prohibit experts from testifying to identification unless they believe themselves certain that they have made a correct match. Experts therefore make only what they term “positive” or ‘absolute’ identifications—essentially making the claim that they have matched the latent print to the one and only person in the entire world whose fingertip could have produced it.”); see also Jonathan J. Koehler, *Fingerprint Error Rates and Proficiency Tests: What They Are and Why They Matter*, 59 HASTINGS L. J. 1077, 1077 (2008).

11. See Cole, *Individualization Is Dead*, *supra* note 9, at 121–22 (citing various TWGFAST and SWGFAST reports); *Scientific Working Group on Friction Ridge Analysis, Study and Technology*, 52 J. FORENSIC IDENTIFICATION 263, 327 (2002).

12. JOSEPH POLSKI ET AL., THE REPORT OF THE INTERNATIONAL ASSOCIATION FOR IDENTIFICATION, STANDARDIZATION II COMMITTEE 8 (2011) (quoting IAI Resolution VII (1979)).

13. *Id.*

Cole and Matt Barno have put it, “many forensic disciplines historically reported in what can reasonably be characterized as a ‘non-statistical’ manner, using verbal formulations that were categorical rather than continuous, reflecting certainty rather than uncertainty.”¹⁴ Cole and Barno’s empirical analysis of actual testimony by forensic analysts published in 2020 found that analysts in the four disciplines they studied—friction ridge (fingerprints), firearms and toolmarks, questioned documents, and shoeprints—still testify almost entirely in non-probabilistic terms.¹⁵ Those findings are consistent with Cole’s empirical analysis of expert testimony transcripts thirteen years earlier, in 2007.¹⁶ The findings are also consistent with empirical work by Bali and colleagues in 2020 that similarly found that forensic analysts report their findings in categorical, not probabilistic, terms.¹⁷

Such implicit or explicit certainty-of-identification testimonial claims rest on several (mostly very unscientific) assumptions.¹⁸ Among those assumptions is that all fingerprints, or bullet striations, or bitemarks, or whatever the relevant pattern, are unique in the world; that the uniqueness and potentially divergent minutiae in the patterns are reliably discernible by the analyst; that any matching points the analyst discerns between the crime scene evidence and the suspect’s sample can be sufficient to make it a certainty they share a common source; and that analysts can flawlessly examine two sets of prints or samples and compare them to all other prints or samples they had seen in the past so that they can say conclusively that no prior sample they had examined shared those identified features.¹⁹ Because assumptions like these have no basis in science, the National Academy of Sciences, in its pathbreaking 2009 report on the state of forensic sciences in the United States, criticized latent print examiners for presenting their findings of “individualization” as fact rather than opinion and for claiming “100%

14. Simon A. Cole & Matt Barno, *Probabilistic Reporting in Criminal Cases in the United States: A Baseline Study*, 60 *SCI. & JUST.* 406, 406 (2020).

15. *Id.* at 412.

16. See Simon A. Cole, *Where the Rubber Meets the Road: Thinking about Expert Evidence as Expert Testimony*, 52 *VILL. L. REV.* 803, 835 (2007).

17. See Agnes S. Bali, Gary Edmond, Kaye N. Ballantyne, Richard I. Kemp & Kristy A. Martire, *Communicating Forensic Science Opinion: An Examination of Expert Reporting Practices*, 60 *SCI. & JUST.* 216, 216 (2020).

18. See William C. Thompson, Joëlle Vuille, Franco Taroni, & Alex Biedermann, *After Uniqueness: The Evolution of Forensic Science Opinion*, 102 *JUDICATURE* 18, 19 (2018).

19. See, e.g., Robert Epstein, *Fingerprints Meet Daubert: The Myth of Fingerprint “Science” Is Revealed*, 75 *S. CAL. L. REV.* 605, 613 (2002).

certain[ty],” the ability to match to the suspect to “exclusion of all others,” and a “zero error rate.”²⁰

Criticisms like these have led most forensic disciplines to soften their prior claims of certainty and to begin to recognize the probabilistic nature of their enterprise.²¹ Analysts in fields including fingerprint comparison are now typically instructed to refrain from claiming certainty, and rather to frame their claims instead in vague terms that, at least if pushed, accept the possibility of error or a coincidental match.²² By 2011, the Standardization II Committee of the IAI, for example, recommended a new guideline providing that “[a]ny member or certified latent print examiner may offer oral or written reports of testimony of probable or likely conclusions concerning source attribution of two friction ridge impressions being from the same source”²³ Today, in the United States, most disciplines instruct their analysts to testify to “identifications” while recognizing those opinions are opinions, not absolute certainties.

Indeed, in their published policy papers many disciplines now recognize the probabilistic nature of their determinations, even if analysts do not always make that clear in their reports and testimony. In 2017, for example, the Friction Ridge Subcommittee of the Organization of Scientific Area Committees (OSAC) of the National Institute of Standards and Technology issued guidelines providing:

In order to reach an identification decision ‘to the exclusion of all others,’ there would need to be an assumption of uniqueness, and the entire world’s population would need to be considered, and rejected, as a potential source of the unknown impression. These two claims are neither supportable, nor necessary, to form an opinion of source identification within a relevant population.²⁴

The OSAC further described the process of fingerprint analysis in ways that make the probabilistic nature of the enterprise apparent:

20. NAS REPORT, *supra* note 4, at 143.

21. Heidi Eldridge, *The Shifting Landscape of Latent Print Testimony: An American Perspective*, 3 J. FORENSIC SCI. & MED. 72, 78 (2017).

22. For a discussion of the competing norms for expressing findings in probabilistic terms, see Thompson et al., *supra* note 18, at 21–26.

23. POLSKI et al., *supra* note 12, at 3.

24. FRICTION RIDGE SUBCOMM., PHYSICS/PATTERN SCI. AREA COMM., ORG. OF SCI. AREA COMMS. (OSAC) FOR FORENSIC SCI., GUIDELINE FOR THE ARTICULATION OF THE DECISION-MAKING PROCESS LEADING TO AN EXPERT OPINION OF SOURCE IDENTIFICATION IN FRICTION RIDGE EXAMINATIONS § 4.8.2.2.1 (2017).

An examiner considers, based upon knowledge and experience, the probability of encountering the observed corresponding features in two impressions made by the same source against the probability of observing the same correspondence between the unknown impression and an impression from a different source. In order to support the proposition that the two impressions were made by the same source, an examiner must find discriminability in the corresponding features to outweigh any support for the proposition that the two impressions were made by different sources. The degree to which support for a proposition of same source outweighs support for a proposition of different source is the strength of the evidence.²⁵

Yet, as some of the language above suggests, OSAC still permits examiners to make categorical, not probabilistic, identification claims in what it acknowledges is a probabilistic field. In its 2017 Report, the Friction Ridge OSAC asserts that “[s]ource identification is the opinion by an examiner that two friction ridge skin impressions originated from the same source. This opinion is the decision that the features are in sufficient correspondence and that the probability the questioned impression was made by a different source is so small that it is negligible.”²⁶ The guidelines explain: “Rather than expressing a source identification as an incontrovertible fact, the friction ridge discipline is now articulating the source identification conclusion as a decision that is expressed as an expert opinion.”²⁷

Such tempered language still presents problems, however. The language now implicitly admits that the determinations are probabilistic, not certain, but to some degree hides that caveat. If not pushed or explored fully on cross-examination, such opinions will sound like categorical, definitive assertions, rather than probabilistic opinions.

Moreover, to the degree that the disciplines acknowledge that the “identifications” made by analysts are probabilistic, those probabilistic opinions implicitly suggest some empirical basis for determining the likelihood, or the probability, that a “match” means the accused was indeed the source of the crime scene evidence. For all but DNA profiling, however, there is almost no data, no statistical calculation—in the end, no science—to support those essentially

25. *Id.*

26. *Id.* at 8.

27. *Id.* at 9.

statistical claims.²⁸ As the President’s Council of Advisors on Science and Technology observed in 2016, categorical claims of “identify,” or a “match,” or “similarity” between samples are scientifically indefensible without underlying data to inform the fact-finder of the significance of those similarities: “Without appropriate estimates of accuracy, an examiner’s statement that two samples are similar—or even indistinguishable—is scientifically meaningless: it has no probative value, and considerable potential for prejudicial impact.”²⁹

This problem with most of the pattern-matching disciplines, however, was invisible to the legal system until DNA came along and exposed the inadequacy of this approach. Even now, it is obscured by standards that permit non-probabilistic claims of certainty as to unavoidably probabilistic comparisons. The justification for asserting certainty? The very absence of data on which to calculate probabilities. If in a sense the claim of “identification” now carries an asterisk, it is not at all clear that lawyers and judges see the asterisk or know what it means. And what it means should negate the claimed certainty.

B. Revelations from DNA Evidence

Unlike the other disciplines, DNA was actually based on scientific study, and built its probabilistic claims on real population data and statistical calculations based on that data. With DNA, analysts could tell from population sampling how frequently one would expect to see particular DNA alleles (the specific combinations and repeats of base pairs—the nucleotides that make up the DNA ladders) in targeted areas, or loci, on a DNA strand.³⁰ From that, they could tell, by multiplying the expected population frequencies of the alleles at each locus by each other, what the odds were that a particular profile would match DNA from a random person in the population—thereby creating a random match probability.³¹ As the number of loci tested has increased, so has the statistical confidence that a random match is not present. It is through that process that DNA analysts are able to say, with real numbers and scientific justification, that the chances that a random person in the popula-

28. Work is beginning now to provide data for statistical claims, such as data on the frequency of particular fingerprint patterns, but such efforts are at a fairly nascent stage for most disciplines. See Thompson et al., *supra* note 18, at 22.

29. PCAST REPORT, *supra* note 1, at 6.

30. See, e.g., William C. Thompson, Simon Ford, Travis Doom, Michael Raymer & Dan Krane, *Part I: Evaluating Forensic DNA Evidence*, CHAMPION, Apr. 2003, at 116, 118.

31. *Id.*

tion would have the same DNA profile as that found at the crime scene is one in hundreds of millions, billions, quintillions, and even more.³² None of the other pattern-matching disciplines at present has the capacity to calculate probabilities based on real numbers, like random match probabilities.³³

This observation is hardly new (although it still has not fully permeated the practitioners' worlds in either law or forensic science). As Michael Saks and Jonathan Koehler observed 15 years ago, the individualization disciplines (i.e., the pattern-matching disciplines) permitted "forensic scientists to draw bold, definitive conclusions that can make or break cases . . ." ³⁴ without "developing measures of objective attributes, collecting population data on the frequencies of variations in those attributes, testing attribute independence, or calculating and explaining the probability that different objects share a common set of observable attributes."³⁵ Saks and Koehler, and others since then, have therefore observed that:

DNA typing can serve as a model for the traditional forensic sciences in three important respects. First, DNA typing technology was an application of knowledge derived from core scientific disciplines. This provided a stable structure for future empirical work on the technology. Second, the courts and scientists scrutinized applications of the technology in individual cases. As a result, early, unscientific practices were rooted out. Third, DNA typing offered data-based, probabilistic assessments of the meaning of evidentiary "matches." This practice represented an advance over potentially misleading match/no-match claims associated with other forensic identification sciences.³⁶

Unfortunately, for virtually all pattern-matching disciplines other than DNA, the "science" has not adequately moved beyond those "misleading match/no-match," or "identification/exclusion," claims. The National Academy of Sciences observed in its 2009 Report that "no forensic method other than nuclear DNA analysis has been rigorously shown to have the capacity to consistently and with a high degree of certainty support conclusions about 'individualiza-

32. See Bruce S. Weir, *The Rarity of DNA Profiles*, 1 ANNALS APPLIED STATS. 358, 358 (2007).

33. See NAS REPORT, *supra* note 4, at 44; PCAST REPORT, *supra* note 1, at 11, 12, 49, 64.

34. Saks & Koehler, *supra* note 3, at 892.

35. *Id.*

36. *Id.* at 893. See also NAS REPORT, *supra* note 4, at 42 ("The increased use of DNA analysis as a more reliable approach to matching crime scene evidence with suspects and victims has resulted in the reevaluation of older cases that retained biological evidence that could be analyzed by DNA.").

tion' (more commonly known as 'matching' of an unknown item of evidence to a specific known source)."³⁷ And part of the reason for this, identified by the NAS, is that "[t]he determination of uniqueness [i.e., 'identification'] requires measurements of object attributes, data collected on the population frequency of variation in these attributes, testing of attribute independence, and calculations of the probability that different objects share a common set of observable attributes"—but most forensic disciplines lack such essential measurements.³⁸ The NAS observed, for example, that "population statistics for fingerprints have not been developed, and friction ridge analysis relies on subjective judgments by the examiner."³⁹ Moreover, as PCAST and others have recognized, an essential component of making valid probabilistic claims is an assessment of error rates, yet none of the disciplines have adequately studied the error rates in their respective fields.⁴⁰ More than eleven years later, the NAS Report's observations still hold.

C. *The Microscopic Hair Analysis Example*

A clear example of the problems with this approach can be seen in the way that forensic analysts typically reported on the results of microscopic hair analysis. In some respects, microscopic hair analysts were more transparent than their colleagues from other pattern-matching disciplines, because hair microscopists were trained that they never could make definitive "matches," or render "identification" opinions, because there could be no claim to uniqueness of an individual's microscopic hair features as would be required to permit such definitive opinions.⁴¹ Ethical hair microscopists, following the rules, would make the limits of their discipline clear in this regard. But hair microscopists still ran into trouble, in ways that are illuminating for other disciplines, by hiding the nature of their probabilistic claims or by overstating the power

37. NAS REPORT, *supra* note 4, at 87.

38. *Id.* at 44.

39. *Id.* at 139.

40. See PCAST REPORT, *supra* note 1, at 33, 53; Jonathan J. Koehler, *Forensics or Fauxrensics? Ascertaining Accuracy in the Forensic Sciences*, 49 ARIZ. STATE L.J. 1369, 1380 (2017) ("[E]ven if a DNA analysis indicates that a DNA profile is so rare there is probably only one person on the planet who could be its source, the reliability of that reported match will ultimately turn on the overall risk of a false positive error . . . and not merely on the small risk that the match is coincidental.").

41. See Cary T. Oien, *Forensic Hair Comparison: Background Information for Interpretation*, 11 FORENSIC SCI. COMM'NS (Apr. 2009), <https://bit.ly/30zytv2> [<https://perma.cc/L5J4-36RH>].

of their opinions by implying the presence of statistics and data that simply did not exist.

In 2015, the FBI joined with the Innocence Project and the National Association of Criminal Defense Lawyers to review the testimony of its hair microscopists to evaluate it for scientific validity.⁴² What this review found was startling: preliminary results showed that at least 90% of the hair microscopists' testimony contained erroneous statements.⁴³

What was the nature of those scientific errors? Often, the errors involved implying statistical power in the analysts' opinions that had no basis in research, data, or statistical analyses. For example, in one fairly typical case the analyst had testified, in part, as follows:

[I]t's my experience and the collective experience of the FBI Laboratory that rarely, extremely rarely do we see known hairs from two different people from the same person that we can't tell them apart. So, when I associate a known hair from a standard in my opinion, it carries a high degree of probability it originated from that person, although I can't say it's a positive means of identification⁴⁴

The analyst also testified:

In approximately 3,000 cases examined in eight and a half years almost all of those cases have at least two known standards, the victim and the suspect hairs. Some of them have many, many as ten or twenty or thirty known hair standards and on only one occasion have I had known hairs from two different people that I compared I could not tell them apart and those were Negroid, but like hairs, I have never had any Caucasian head hairs, I have never not been able to differentiate two Caucasian even head hairs.⁴⁵

The analyst then rendered his opinion:

I compared this [crime scene] hair with the known head hair standards from the defendant . . . [and it] was microscopically the same as the known head hairs of the defendant. I concluded that

42. Meterko, *supra* note 3, at 643 (citing *FBI Testimony on Microscopic Hair Analysis Contained Errors in at Least 90 Percent of Cases in Ongoing Review*, FBI (Apr. 20, 2015), <https://bit.ly/3esQQKm> [<https://perma.cc/K8P3-K4Q6>]).

43. *Id.*

44. Trial Transcript on file with author. The case name is omitted here to protect client confidentiality.

45. *Id.* at 181-82.

this head hair found [at the scene] was consistent with originating from the defendant.⁴⁶

To bolster his opinion, the analyst then testified that a more senior analyst had confirmed his conclusions, and that analyst had “worked in excess of ten thousand cases”⁴⁷

The FBI subsequently recognized that this testimony was scientifically invalid and wrote a letter to the prosecutor in the case alerting him to the errors. In essence, the analyst had made probabilistic and statistical claims that had no basis. The analyst had no data to support his claim that when he makes an association it means there is a high degree of probability that the crime scene hair originated from the accused.⁴⁸ Indeed, subsequent mitochondrial DNA testing conclusively established that the crime scene hair was not the defendant’s hair; it was in fact, not a “match,” and the analyst’s association was incorrect. Moreover, to the extent the analyst gave numbers, suggesting that he might have an error rate of less than 1 in 3,000, or that his senior colleague had an even lower error rate perhaps approaching less than 1 in 10,000, the claims were nonsensical.⁴⁹ The analyst’s own assessment that in examining 3,000 hairs he had only once failed to distinguish like hairs is purely a

46. *Id.* at 60-61.

47. *Id.* at 63-65.

48. The FBI acknowledged, for example:

The examiner assigned to the positive association a statistical weight or probability or provided a likelihood that the questioned hair originated from a particular source, or an opinion as to the likelihood or rareness of the positive association that could lead the jury to believe that valid statistical weight can be assigned to a microscopic hair association. This type of testimony exceeds the limits of the science.

Postconviction Hearing Transcript, on file with the author, at 139.

49. Hence, the FBI also acknowledged:

The examiner cites the number of cases or hair analyses worked in the lab and the number of samples from different individuals that could not be distinguished from one another as a predictive value to bolster the conclusion that a hair belongs to a specific individual. This type of testimony exceeds the limits of the science.

Postconviction Transcript, *supra* note 48, at 145-46.

A defense expert in postconviction proceedings explained the error in this way:

And the reason . . . why it exceeds the limit of science is because there’s an implication there that the analyst is taking every hair they look at in a case, a questioned hair, and somehow comparing it back to these 3,000 hairs or 30,000 hairs or whatever it is that they’ve looked at previously. And of course that’s impossible to do. Why? Because, as I said before, they’re not taking this hair and looking at getting the slides from each case and looking back at those. That’s the only way you could remember. You cannot remember these characteristics.

Id. at 147. He added:

subjective assessment of the type that PCAST warned against—”[w]ithout appropriate estimates of accuracy, an examiner’s statement that two samples are similar—or even indistinguishable—is scientifically meaningless: it has no probative value, and considerable potential for prejudicial impact.”⁵⁰ In this regard, PCAST’s elaboration is also germane:

We note, finally, that neither experience, nor judgment, nor good professional practices (such as certification programs and accreditation programs, standardized protocols, proficiency testing, and codes of ethics) can substitute for actual evidence of foundational validity and reliability. The frequency with which a particular pattern or set of features will be observed in different samples, which is an essential element in drawing conclusions, is not a matter of “judgment.” It is an empirical matter for which only empirical evidence is relevant. Similarly, an expert’s expression of *confidence* based on personal professional experience or expressions of *consensus* among practitioners about the accuracy of their field is no substitute for error rates estimated from relevant studies.⁵¹

While microscopic hair analysis is rarely offered as evidence in criminal cases any longer, primarily because mitochondrial DNA testing on hair shafts is far more reliable,⁵² the example of misleading probabilistic or statistical testimony remains instructive because the same type of testimony is routinely offered in other pattern-matching disciplines today.⁵³ The forensic science community just has not yet recognized the problematic nature of such claims as widely in the other disciplines.

D. *The Shaken Baby Syndrome/Abusive Head Trauma Example*

The problem of misleading or hidden, yet baseless, probabilistic claims is not limited to the pattern-matching disciplines. Prominent among other forms of forensic expert opinions in this regard is medical expert opinion evidence on child abuse, and in particular

But to say that all the experience of all the hairs that the agent or other examiner has looked at over the years in individual cases are somehow being compared back to this one hair that’s being looked at in this particular case, this one questioned hair, is just ludicrous, because it can’t be done.

Id. at 149.

50. PCAST REPORT, *supra* note 1, at 6.

51. *Id.*

52. *Microscopic Hair Comparison*, CAL. INNOCENCE PROJECT, <https://bit.ly/3t6C4wW> [<https://perma.cc/K2GU-ZU7R>].

53. *See, e.g., supra* notes 25–27 and accompanying text.

Shaken Baby Syndrome (SBS), now known more broadly as Abusive Head Trauma (AHT).⁵⁴ The science underlying SBS/AHT has become increasingly controversial in recent years, given that it is inherently difficult to conduct high-quality studies on the effects of shaking or striking the head of an infant on a hard surface.⁵⁵ Because of these research challenges, virtually all knowledgeable researchers and physicians agree that there is no definitive test for abuse, and hence all medical opinions are inherently probabilistic.⁵⁶ As one leading group of proponents of the SBS/AHT hypothesis put it: “Gold standard definitional criteria for AHT do not exist.”⁵⁷ Accordingly, these proponents acknowledged, “in the absence of a gold standard, clinicians can rarely confirm or exclude AHT with complete certainty and are compelled instead to adopt a probabilistic approach to the diagnosis.”⁵⁸

Yet, as with the pattern-matching disciplines, expert medical testimony on SBS/AHT rarely addresses the probabilistic nature of the expert’s opinions. Physicians opining that a child was a victim of violent shaking or shaking with impact routinely offer those opinions “to a reasonable degree of medical certainty”—a term that itself is meaningless in both science and law and hence inherently misleading⁵⁹—as definitive diagnoses.⁶⁰ Even when more equivocal and obviously probabilistic language is used, however, us-

54. For a discussion about why opinions about SBS/AHT are not diagnoses in the true sense, but rather are etiological determinations, see Keith A. Findley, D. Michael Risinger, Patrick D. Barnes, Julie A. Mack, David A. Moran, Barry C. Scheck & Thomas L. Bohan, *Feigned Consensus: Usurping the Law in Shaken Baby Syndrome/Abusive Head Trauma Prosecutions*, 2019 WIS. L. REV. 1211, 1238-45 [hereinafter Findley et al., *Feigned Consensus*].

55. *See id.* at 1244.

56. *See* Keith A. Findley, Patrick D. Barnes, David A. Moran & Waney Squier, *Shaken Baby Syndrome, Abusive Head Trauma, and Actual Innocence: Getting It Right*, 12 HOUS. J. HEALTH L. & POL’Y 209, 282, 312 (2012) [hereinafter Findley et al., *Getting It Right*].

57. Kent P. Hymel et al., *Derivation of a Clinical Prediction Rule for Pediatric Abusive Head Trauma*, 14 PEDIATRIC CRITICAL CARE MED. 210, 212 (2013).

58. *Id.* at 217.

59. NAT’L COMM’N ON FORENSIC SCI., USE OF THE TERM “REASONABLE DEGREE OF SCIENTIFIC CERTAINTY” 1 (2016), <https://bit.ly/3qJ0Kdp> [<https://perma.cc/73DX-VT5S>] (recommending that the legal community not ask experts to testify to a reasonable degree of . . . certainty, “[as such terms] have no scientific meaning and may mislead [jurors or judges] when deciding whether guilt has been proved beyond a reasonable doubt”); Paul C. Gianelli, “Reasonable Scientific Certainty”: *A Phrase in Search of a Meaning*, 25 CRIM. JUST. 40, 41 (“The term ‘reasonable medical certainty’ has no scientific meaning. Its legal meaning is at best ambiguous, at worst misleading.”). Note specifically that, where SBS/AHT has no data by which to assess probabilities precisely, certainty necessarily is impossible—which makes the term “medical certainty” contradictory and its preface, “reasonable degree,” empty.

ing terms such as “concern for,” “consistent with,” “suspicious for,” “suggestive of,” “highly associated with,” and the like, physicians typically go on to render an opinion about a “diagnosis”—a determination that the child was violently shaken or both shaken and slammed. As Dr. Steven Gabaeff has observed, however, these equivocal terms “connote possibility, plausibility, and probability They should not be used to connote legal or medical certainty, or a final diagnosis, simply because they do not.”⁶¹ Yet these terms frequently lead to conclusions expressed as a final diagnosis.

Moreover, as with the pattern-matching disciplines, those probabilistic assessments are not based on any reliable statistical data about the actual odds of abuse given any presenting medical findings, or any assessment of error rates in diagnosis in general or in this particular “diagnosis” (because, for this particular “diagnosis,” in the absence of a gold-standard diagnostic standard or test for abuse, there is no way to calculate an error rate). I return to a deeper discussion of the problems with use of (or failure to use) statistics in SBS/AHT determinations in the next two sections of this essay.

II. MAKING OVERTLY STATISTICAL CLAIMS, BUT BOTCHING THE NUMBERS: MORE ON THE SBS/AHT EXAMPLE

When experts do acknowledge and try to draw on statistical data, they sometimes just get the statistics wrong. The full range of statistical errors that analysts and lawyers make when dealing with forensic science evidence is beyond the scope of what I can cover here. To illustrate the extent and nature of just some of these errors, I will draw on statistical claims made in one particular type of forensic expertise: the medical opinion of child abuse, again and in particular, SBS/AHT. Here, the statistical errors are numerous; I address only a few by way of illustration.

A. *Miscalculating the Probabilities*

To start, as noted above, physicians routinely make probabilistic determinations of abuse even though they have no reliable data upon which to base those probabilistic assessments. It is no answer to say that physicians rely upon clinical judgment, because to be valid that clinical judgment must be based on something reliable—

60. See Steven C. Gabaeff, *Recognizing the Misuse of Probabilistic Language and False Certainty in False Accusations of Child Abuse*, 4 J. RSCH. PHIL. & HIST. 1, 4 (2021), available at <https://bit.ly/3bFqvqG> [<https://perma.cc/BS4M-MDPV>].

61. *Id.* at 5.

preferably data, and if not data, at least on feedback that enables learning from experience. But because a “diagnosis” of abuse can never be assessed for accuracy by independent criteria—again, there is no gold standard test for abuse—and because there is also no treatment a physician can prescribe for abuse to follow and assess whether the determination was correct, a physician might simply be making the same mis-determination of abuse for years, with no opportunity to learn from that experience or those mistakes.⁶²

To the extent that physicians attempt to ascribe some diagnostic power to the presence of particular medical findings (e.g., subdural hematoma, retinal hemorrhages, cerebral edema, bone fractures, and others), they tend to do so in ways that fail to take into account all of the relevant factors needed to determine the diagnostic power of those findings. And on the factors they do consider, they often make fundamental analytical errors. In the following subsections I take up each of the necessary considerations for a valid statistical or probabilistic assessment of the likelihood that a child was abused, when that assessment is based on the presenting medical findings in a case.

1. *Overlooking the Base Rate*

As a starting point, a proper assessment of the probability of abuse given a set of medical findings (again, findings like subdural hematoma, retinal hemorrhages, cerebral edema, bone fractures, or the like), requires consideration of the base rate—that is, the frequency with which abuse occurs in a population.⁶³ Such base rates are notoriously difficult to estimate, but they nonetheless must be factored into any attempt to ascertain the probability that a particular child was abused. The population of interest might be as broad as “children in the United States between the ages of X and Y,” or somewhat narrower, such as “toddlers in foster homes.” Although there are no firm rules for identifying the populations from which to compute base rates, because a majority of children in most identifiable populations are probably not abused, the base rate will tend to suggest that a child has not been abused. Of course, specific

62. For a fuller discussion of the problem of absent feedback in medical determinations of abuse, see Findley et al., *Feigned Consensus*, *supra* note 54, at 1243–44 (e.g., “Because opining about the etiology of a child’s brain findings provides no feedback mechanism, the entire enterprise is untethered from empirical confirmation. Without the feedback required to ‘engage in learning,’ the expert’s opinions based on clinical judgment can amount to little more than *ipse dixit*, which the Supreme Court has recognized as problematic under the Federal Rules of Evidence.”).

63. Findley et al., *Getting It Right*, *supra* note 56, at 287–90.

medical evidence may suggest otherwise, and this evidence could be more statistically compelling in the direction of abuse than the non-abuse hypothesis favored by the base rate. But the base rate must nonetheless be considered as a starting point, for reasons that are explained below. Yet, as we shall see, child abuse physicians routinely ignore the base rate and opine about child abuse based largely on medical evidence, as if that were all that is needed.

2. *Misunderstanding the Diagnosticity of the Medical Evidence*

To make a proper probabilistic determination, the base rate must then be combined with some measure of the strength of the medical evidence. The strength of the medical evidence in abuse cases is best assessed by comparing the frequency with which a symptom of interest appears in abuse cases in relation to non-abuse cases. This latter frequency can be assessed in a rather straightforward and intuitive manner, employing what Thomas Lyon and Jonathan Koehler have called the “relevance ratio.”⁶⁴ Lyon and Koehler explain that the relevance ratio requires consideration of two proportions:

the proportion of abuse cases in which the symptom occurs, and the proportion of nonabuse cases in which the symptom occurs. If the proportion of abuse cases exhibiting the symptom is greater than the proportion of nonabuse cases exhibiting the symptom, then the symptom is relevant for proving that abuse occurred.⁶⁵

In this ratio, then, the proportion of abuse cases that have the symptom of interest is divided by the proportion of nonabuse cases that also have that symptom. Any resulting number greater than one logically provides at least some evidence supporting a determination of abuse, while any number less than one provides some evidence supporting a determination of nonabuse.

Application of this ratio can reveal some surprising and sometimes counterintuitive, but logically correct, conclusions. Understood properly, for example, this ratio shows that even some symptoms or findings that are extraordinarily frequent in abuse cases might nonetheless be irrelevant to a finding of abuse, if they are also quite frequent in nonabuse cases (e.g., subdural hemato-

64. Thomas D. Lyon & Jonathan J. Koehler, *The Relevance Ratio: Evaluating the Probative Value of Expert Testimony in Child Sexual Abuse Cases*, 82 CORNELL L. REV. 43, 45 (1996).

65. *Id.* at 46.

mas).⁶⁶ Conversely, even some symptoms or findings that are relatively rare in abuse cases may nonetheless be relevant to a finding of abuse, if they are even significantly rarer in nonabuse cases.⁶⁷ Because physicians so frequently fail to consider the logical import of this ratio, however, “the terms experts use to describe the frequency or severity of symptoms often obscure the relevance of those symptoms.”⁶⁸ Commonplace testimony, such as assertions that a finding is “consistent with” abuse, can be very misleading, because such testimony is likely to be understood by factfinders as strong evidence of abuse, when in fact it might actually suggest the contrary—if it is also “consistent with” (i.e., also found in) nonabuse cases, especially if it is even more common in nonabuse cases.⁶⁹

3. *Misunderstanding How to Combine Base Rates and Relevance Ratios*

A formal way to ensure that both the base rate of abuse and the relevance ratio for available medical information are incorporated into a final judgment about abuse in a given case is through the use of Bayes’s Theorem. Bayes’s Theorem is the principle credited to the 17th Century cleric and mathematician Thomas Bayes, which teaches that, logically, the probability that an event or proposition is true is assessed by taking the prior odds of the proposition (one’s pre-existing assessment of the proposition, based on previously considered evidence or assumptions, which can include the base rate) and updating those odds by multiplying the prior odds by the impact of new or additional evidence (typically computed as a “likelihood ratio”) to arrive at a new assessment, known as the “posterior odds.”⁷⁰ The multiplier in this equation—the likelihood ratio—is simply the relationship of the probability of one hypothesis or conclusion divided by its opposite, or more precisely, the relationship of the probability of the observed data under one

66. *Id.* at 58.

67. *Id.*

68. *Id.* at 49.

69. *Id.* at 51 (emphasis in original) (“Typically, the ‘consistent with’ terminology is merely an observation that *at least some* abused children exhibit the condition. Thus, ‘consistent with’ testimony informs a factfinder that the numerator of the relevance ratio is nonzero, but says nothing about the denominator.”).

70. See Richard O. Lempert, *Modeling Relevance*, 75 MICH. L. REV. 1021, 1023–25 (1977).

hypothesis divided by the probability of the data under the opposite hypothesis.⁷¹

Likelihood ratios in forensic medicine can be understood as the probability of seeing the evidence (e.g., the presenting medical findings, such as subdural hematoma or retinal hemorrhages) if the child was abused (the prosecutor's hypothesis) divided by the probability of seeing the evidence if the child was not abused and the conditions were caused by something other than abuse (the defense hypothesis).⁷² Readers will note that the likelihood ratio is essentially the same as Lyon and Koehler's relevance ratio, discussed above.⁷³ In statistical notation, the formula for the likelihood ratio is

$$\frac{\Pr(E|A)}{\Pr(E|\bar{A})}$$

where Pr means probability, E denotes the evidence, A means the medical condition was caused by abuse, and \bar{A} means the condition was not caused by abuse.⁷⁴

But the likelihood ratio is just one part of what a valid statistical assessment requires; also needed is a base rate of abuse to serve as the prior odds of abuse in Bayes's Theorem. To put the base rate (prior odds) and likelihood ratio (significance of the medical findings) into Bayes's Theorem, then, the formula (prior odds x likelihood ratio = posterior odds) is:

$$\frac{\Pr(A)}{\Pr(\bar{A})} \times \frac{\Pr(E|A)}{\Pr(E|\bar{A})} = \frac{\Pr(A|E)}{\Pr(\bar{A}|E)}$$

The point here is that to accurately estimate the probability of abuse, in a case in which a child presents with medical findings that *could be* caused by abuse, one needs to know several things to plug into this equation, including the prior odds of abuse and the significance of the medical findings in relation to abuse (for determining the likelihood ratio, or the relevance ratio). As noted, there is little reliable data on which to compute a likelihood ratio for SBS/AHT. But even if the likelihood ratio could be estimated accurately, child abuse physicians routinely overlook the importance of assessing the

71. As this suggests, to avoid what is known as the "prosecutor's fallacy" it is essential to remember that this analysis requires consideration of the likelihood of seeing the *evidence* if the hypothesis of guilt is true, and *not* the inverse—the likelihood of guilt if given the evidence in question. For a fuller explanation of the prosecutor's fallacy, see *infra* note 78 and accompanying text.

72. For an explanation of Bayes's Theorem, see James M. Wood, *Weighing Evidence in Sexual Abuse Evaluations: An Introduction to Bayes's Theorem*, 1 CHILD MALTREATMENT 25, 26 (1996).

73. Lyon & Koehler, *supra* note 64, at 48.

74. Lempert, *supra* note 70, at 1025.

prior probability of abuse. In this instance, the prior probability of abuse is the base rate of abuse—that is, the rate at which children are abused in the general population. As Dr. Gabaeff has pointed out, the base rate for abuse is quite low.⁷⁵ Hence, even a relatively high likelihood ratio, when multiplied by a low base rate (prior odds), might still provide relatively low posterior odds.⁷⁶ Yet most physicians make their probabilistic, clinical-judgment-based “diagnoses” of SBS/AHT without any reference to, or apparent consideration of, the low base rate.⁷⁷ This is not to say that medical opinions cannot play a role in determining whether a child has been abused, but it is to say that, especially given the current state of research and data in this field, medical opinions based on nothing more than a constellation of non-specific medical findings inherently have only modest or at best unknown probative value. Yet the testimony is frequently presented in ways that conceal that inherent uncertainty and low probative value. Given these uncertainties, other evidence beyond expert opinions (“diagnoses”) is essential to prove guilt beyond a reasonable doubt, however that reasonable doubt standard itself might be quantified.

A further note about this analysis is necessary to avoid a logical error so common in criminal cases it is known as the “prosecutor’s fallacy.”⁷⁸ As the relationship between the various components of Bayes’s Theorem suggests, it is important to be clear that the competing hypotheses at issue when computing a likelihood ratio (or relevance ratio) is, first, the likelihood of seeing the observed evidence (e.g., the relevant medical findings in an SBS/AHT case, or the particular patterns, such as fingerprints, in a pattern-matching case, or a particular DNA profile) if the defendant is guilty, versus, second, the probability of seeing the observed evidence (e.g., the medical findings or the particular patterns or DNA profile) if the defendant is not guilty. Importantly, the probability of *seeing the evidence* if the defendant is guilty must not be confused

75. Gabaeff, *supra* note 60, at 7.

76. *Id.*

77. See Findley et al., *Getting It Right*, *supra* note 56, at 287–90.

78. For a fuller description of the Prosecutor’s Fallacy see *id.* at 287–90. See also William C. Thompson & Edward L. Schumann, *Interpretation of Statistical Evidence in Criminal Trials: The Prosecutor’s Fallacy and the Defense Attorney’s Fallacy*, 11 L. & HUM. BEHAV. 167, 170–71, 181–82 (1987); Michael I. Meyerson & William Meyerson, *Significant Statistics: The Unwitting Policy Making of Mathematically Ignorant Judges*, 37 PEPP. L. REV. 771, 778 (2010) (footnotes omitted) (“[T]he ‘prosecutor’s fallacy’ . . . incorrectly reverses events in a conditional probability to create a direct statement about the defendant’s probability of guilt that is not implied by the evidence. In logical reasoning, such an error is called ‘transposing the conditional.’”).

with the probability that *the defendant is guilty* given the observed evidence. The likelihood of guilt is an expression of the posterior odds, which requires combining the likelihood ratio with other evidence in the case, including base rates (prior odds) and other information considered by the fact finder. In the same way, the probability of guilt given the evidence is obtained by combining the probability of observing the evidence if the defendant is guilty with the same base rates and prior information. The confusion arises when the two probabilities are treated as the same thing—that is, the prosecutor’s fallacy arises from taking the odds of seeing the evidence if the defendant is guilty and transposing that into a statement of the likelihood that the defendant is in fact guilty. The fallacy would be present, for example, if one were to assert that, because the probability of a random match to a particular DNA profile is one in a billion, then the odds that the defendant who shares that DNA profile with crime scene evidence is guilty are a billion to one. But of course, despite the long odds of a random match, the true odds of guilt (the posterior odds) might be quite small, even close to zero, *depending on the other evidence in the case*—if for example, other evidence shows incontrovertibly that the defendant was in prison at the time of the crime or otherwise was likely not the perpetrator.

An example of the failure to apply Bayes’s Theorem correctly can even be seen in much of the research literature on SBS/AHT. My colleagues and I have previously pointed this out, explaining:

In these Studies [relied upon by child abuse physicians], the correlation of subdural hematoma to abuse is very high but the base rate of abuse compared to non-abuse—to the extent it is revealed in the studies—is sometimes relatively modest, suggesting that subdural hematomas are at best only weakly diagnostic of abuse. Bechtel et al., for example, studied 82 children admitted for head trauma and concluded that 15 (18%) of the injuries were inflicted and 67 (82%) were “accidental.”^[79] Bechtel then reported that 80% (12/15) of the “inflicted” group had subdural hematomas while only 27% (18/67) in the “accidental” group had subdural hematomas.^[80] From this, [child abuse pediatrician] Dr. [Sandeep] Narang concludes that, with a P-value of .001, “the association of SDHs with inflicted injury was highly statistically sig-

79. Kirsten Bechtel et al., *Characteristics that Distinguish Accidental from Abusive Head Trauma in Hospitalized Young Children with Head Trauma*, 114 PEDIATRICS 165, 165 (2004).

80. *Id.* at 167.

nificant.”^[81] But that is only part of the story. When one factors in the low base rate of abuse, the conclusion is quite different. To compute the posterior probability of abuse, which more accurately reflects the diagnostic significance of subdural hematoma, one has to multiply the base rate by the likelihood ratio, which represents “the relative probability of coming across a particular piece of evidence in one group rather than in another.”^[82] Here, since 80% of purported inflicted cases have subdural hematomas and 27% of accidental cases have subdural hematomas, the likelihood ratio is 80:27, or 2.96:1. But because the base rate of abuse is only 18%, the true likelihood of abuse given subdural hematoma is only 35% One can make the same calculation in a different manner: since 18 of the subdural hematomas identified by Bechtel were accidental and 12 were inflicted, subdural hematomas were 50% more common in accident cases than in abuse cases. Either way, subdural hematoma is not diagnostic of abuse since most cases with this finding are nonabusive

A similar analysis applies to other studies. In the Matschke study, for example, the authors looked at 715 infant deaths, finding subdural hematomas in 50 of them.^[83] Unlike the Bechtel study, the Matschke study attempted to identify all causes of the subdural hematomas, not just those attributed to trauma. Of the 50 cases with subdural hemorrhage, 15 (30%) were identified as traumatic and 35 (70%) were attributed to other causes.^[84] Of the 35 cases that were not identified as traumatic, the subdural hemorrhages were attributed to bleeding/clotting disorders, perinatal events, infections, metabolic diseases, or (in 8% of the cases) undetermined causes.^[85] A simple counting reveals that the study does not support the conclusion of its authors, which Dr. Narang quotes for the proposition that “most SDH’s are attributable to trauma.”^[86] To the contrary, the data show that most SDHs are attributable to non-traumatic events, by a ratio of 70% to 30% As this suggests, while Dr. Narang is undoubtedly correct that some children who have been abused will have subdural hemorrhages, [he errs] when he claims that children

81. Sandeep Narang, *A Daubert Analysis of Abusive Head Trauma/Shaken Baby Syndrome*, 11 HOUS. J. HEALTH L. & POL’Y 505, 545 (2011).

82. Wood, *supra* note 72, at 26.

83. Jakob Matschke et al., *Nonaccidental Head Injury is the Most Common Cause of Subdural Bleeding in Infants <1 Year of Age*, 124 PEDIATRICS 1587, 1587 (2009).

84. *Id.*

85. *Id.* at 1589.

86. Narang, *supra* note 81, at 543 (citing Matschke et al., *supra* note 83, at 1594).

who have subdural hemorrhages are likely to have been abused. Instead, this is just one of many possible causes.⁸⁷

Much of the research on SBS/AHT is afflicted with this failure to attend to the base rate and the likelihood ratio (i.e., the relevance ratio), and hence it claims high diagnostic power (probative value) of various medical findings when there might actually be little or none.

B. *Miscalculating the Rarity of Alternative Causes*

When child abuse physicians consider possible alternative causes of the medical findings used to identify SBS/AHT, they routinely stumble over other statistical errors as well. One example is the way physicians consider and then reject some alternative explanations for the relevant medical findings. One of the most common explanations provided by caregivers for an infant or child's brain injuries is a short fall—from a changing table, down a flight of stairs, from a couch or bed, from playground equipment, or the like.⁸⁸ Child abuse physicians, however, routinely disregard those explanations as so unlikely as to be categorically untrue; indeed, they will take a caregiver's explanation of a short fall as a “discrepant history,” which rather than pointing toward innocence, is considered additional evidence of guilt, because it reveals an attempt to hide what really happened.⁸⁹

To justify this disregard for short falls, child abuse physicians point to several studies that claim to establish as a statistical matter

87. Findley et al., *Getting It Right*, *supra* note 56, at 288–90.

88. See Sandeep K. Narang, Amanda Fingarson, James Lukefahr & the Council on Child Abuse & Neglect, *Abusive Head Trauma in Infants and Children*, 145 PEDIATRICS 1, 2 (2020) (“Short falls (often defined as less than 1.5 m, or 5 ft) continue to be a common historical explanation for injuries often seen in AHT.”).

89. See, e.g., Ann-Christine Duhaime et al., *Head Injury in Very Young Children: Mechanisms, Injury Types, and Ophthalmologic Findings in 100 Hospitalized Patients Younger than 2 Years of Age*, 90 PEDIATRICS 179, 184 (1992) (stating that “[m]ost determinations of nonaccidental injury are based on the notion of ‘history insufficient to explain injuries,’” and contending that child abuse should be suspected or presumed when a child suffers intracranial injury and the caregiver offers a short fall as the precipitating event); Findley et al., *Getting It Right*, *supra* note 56, at 265 (“If the parent or caretaker . . . describes a short fall or no trauma at all, the history is deemed to be inconsistent with the [medical] findings, and the case is classified as abusive.”); *id* at 248 n.139 (“This is another example of the circularity that affects much of the research in this field. If deaths presenting with [supposedly tell-tale brain injuries] following a reported short fall are typically diagnosed as SBS/AHT, the number of accidental short fall fatalities will appear to be vanishingly small. The rarity of short fall fatalities is then used to reject the caretaker’s history of a short fall and to support an SBS/AHT diagnosis.”).

that short falls almost never, if ever, result in death.⁹⁰ The first group of these studies examined witnessed short falls in controlled environments, such as falls (or drops of infants) in hospitals. One study reported on 207 children who fell out of bed in the hospital, none of whom died.⁹¹ Those results were consistent with other similar studies, with all the studies combined providing a total of approximately 600 observed short falls, none of which resulted in death.⁹² But to conclude from these studies—as child abuse pediatricians have done repeatedly in testimony—that children do not die from short falls, is to seriously misunderstand the data. For a rare event—and everyone agrees that death from a short fall is a rare event⁹³—one would not expect to see short falls show up in such small sample sizes. If one were to assume that death resulted from a short fall even as frequently as 1 time in 1000 times (and no one contends short-fall deaths are that common), it would not be at all surprising that no children died from falls in sample sizes of 207, or even 600, or more. To suggest otherwise is akin to standing on a street corner, observing a few hundred cars go by without a collision, and taking that as proof that cars do not crash. As Leila Schneps has observed, for a clinician to reach a conclusion that short falls do not cause death “he or she needs to be certain that such a thing is not merely rare, but has actually never occurred. Such a claim must necessarily be based on large-scale studies, since studies of just a few individuals are unlikely to reveal rare events.”⁹⁴

90. See David L. Chadwick et al., *Deaths from Falls in Children: How Far is Fatal?*, 31 J. TRAUMA 1353, 1355 (1991); David L. Chadwick et al., *Annual Risk of Death Resulting from Short Falls Among Young Children: Less Than 1 in 1 Million*, 121 PEDIATRICS 1213, 1213 (2008); R. Helfer et al., *Injuries Resulting When Small Children Fall Out of Bed*, 60 PEDIATRICS 533 (1977).

91. Thomas J. Lyons & R. Kim Oates, *Falling Out of Bed: A Relatively Benign Occurrence*, 92 PEDIATRICS 125, 126 (1993).

92. Pravit Nimityongskul & Lewis Anderson, *The Likelihood of Injuries When Children Fall Out of Bed*, 7 J. PEDIATRIC ORTHOPEDICS 184, 184-86 (1987); S. Levene & G. Bonfield, *Accidents on Hospital Wards*, 66 ARCHIVES DISEASE CHILDHOOD 1047, 1047-49 (1991); S. Monson et al., *In-Hospital Falls of Newborn Infants: Data from a Multihospital Healthcare System*, 122 PEDIATRICS 227, 278-80 (2008); C. Ruddick et al., *Head Trauma Outcomes of Verifiable Falls in Newborn Babies*, 95 ARCHIVES DISEASE CHILDHOOD FETAL NEONATAL ED. 144, 144-45 (2010); Patricia L. Schaffer et al., *Pediatric Inpatient Falls and Injuries: A Descriptive Analysis of Risk Factors*, 17 J. SPECIALISTS PEDIATRIC NURSING 10 (2012).

93. See Findley et al., *Feigned Consensus*, *supra* note 54, at 1230 (“[E]veryone agrees that short-fall deaths are rare.”).

94. Leila Schneps, *When Lack of Information Leads to Apparent Paradoxes and Wrong Conclusions: Analysis of a Seminal Article on Short Falls 2* (2021) (draft manuscript on file with the author) [hereinafter Schneps, *Paradoxes & Wrong Conclusions*].

Even more frequently cited are a pair of papers by Dr. David Chadwick and his colleagues that attempt to calculate a rate of death from short falls. The first, published in 1991, examined data collected over 42 months on children who had sustained injuries in a fall and who were brought to the Children's Trauma Center in San Diego, California.⁹⁵ Of the injured children, the data provided details on the height of the fall in 283 cases. The data showed, surprisingly, that children who reportedly fell the shortest distance (1–4 feet) had the highest fatality rate (7 fatalities out of 100 cases) compared to medium-height falls (5–9 feet, 0 fatalities out of 65 cases) and high falls (greater than 9 feet, 1 fatality out of 118 cases).⁹⁶ From this, Chadwick et al. surmised that the claimed short-fall deaths must be false accounts, because it is simply illogical to believe what this data appears to suggest, that short falls are more lethal than high falls. Chadwick wrote, "The best explanation of the findings is that for the seven children who died following short falls the history was falsified."⁹⁷ Influential child abuse pediatrician Sandeep Narang and his colleagues subsequently reported Chadwick's data and conclusion as a basis for discounting reported short falls as a likely cause of death in these cases, stating, "[i]f reports of deaths from uncorroborated short falls are accepted as valid, then short falls appear to be more dangerous than longer falls."⁹⁸

Chadwick's paper, however, suffers from numerous statistical errors and omissions that lead to unwarranted conclusions. Schneps has pointed out that the first is a problem of missing information—an analytical flaw known as "Simpson's paradox," which arises when missing data "could radically change if not even actually invert a conclusion that appeared obvious from the given data."⁹⁹ Chadwick's data did not report, for example, the ages of most of the children involved in the falls, or of the children who died, yet the ages might be critical in understanding the results.¹⁰⁰ Schneps points out that, generally, "babies are more fragile than toddlers, toddlers more fragile than schoolchildren, and children more fragile than adults."¹⁰¹ But children who fall from great

95. Chadwick et al., *Deaths from Falls in Children*, *supra* note 90, at 1353.

96. *Id.* at 1354.

97. *Id.* at 1355.

98. Sandeep K. Narang et al., *A Daubert Analysis of Abusive Head Trauma/Shaken Baby Syndrome—Part II: An Examination of the Differential Diagnosis*, 13 HOUS. J. HEALTH L. & POL'Y 203, 219 (2013) [hereinafter Narang et al., *Part II*].

99. Leila Schneps, *Short Falls and Shaken Baby Syndrome: Numerical and Reasoning Errors*, at 2 (draft manuscript on file with the author) [hereinafter Schneps, *Numerical & Reasoning Errors*].

100. Schneps, *Paradoxes & Wrong Conclusions*, *supra* note 94, at 3-5.

101. *Id.* at 3-4.

heights must all be older children, as infants and toddlers are generally incapable of climbing to heights. Thus, if data on the children's ages were available, it could very well lead to a conclusion that makes perfect sense out of the death rates in each height category. If indeed the data on ages reflected that infants were disproportionately represented in the short falls, and the high falls were predominantly if not exclusively school children or older, then

[t]he most natural conclusion would seem to be: "Whether short or long, falls are much more dangerous and likely to be fatal in babies up to around age 1 than in older children. The reason there are so few fatalities in the long fall group is because most babies are more supervised and also unable to climb."¹⁰²

Without data on the children's ages, we cannot know if this conclusion is correct, but we also certainly cannot know if Chadwick and Narang's conclusion—that the short-fall death reports must be false—is correct. Failure to include essential information—and the failure to recognize the significance of that information—is a serious error in Chadwick's statistical analysis that undermines its utility.

Second, Chadwick's paper fails to take into account the base rate for short falls, that is, the rate at which children suffer short falls, especially as compared to falls from heights. This is another example of the failure to assess prior odds, as is required for proper analysis under Bayes's Theorem.¹⁰³ To understand why this matters, note that Chadwick reports short-fall and long-fall deaths as if they are roughly equally prevalent occurrences (he reports on 100 short falls, 65 medium-height falls, and 118 high falls), that is, as if they have the same base rate (prior odds).¹⁰⁴ But, obviously, children take short tumbles far more frequently than they fall from heights. Most of those short falls, however, never result in a trip to the hospital, and hence do not represent a recordable event, while virtually all falls from heights do warrant a trip to the hospital, and hence get recorded. Therefore, the number of short fall deaths in Chadwick's analysis cannot be understood in relation to the number of *recorded* short falls in his dataset, but rather only as a proportion of the vastly larger group of unrecorded short falls.

Schneps hypothesizes the difference it might make if the base rate for short falls and long falls were included in the analysis. She writes,

102. *Id.* at 5.

103. See discussion of Bayes's Theorem, *supra* Section II.A.3.

104. *See id.*

Long falls in children are very rare. Every child who takes a long fall of more than 10 feet is brought to the hospital, and all of them are in Chadwick's data set. So we can conclude that there were a total of about 118 long falls over 3.5 years among the children under 15 of San Diego country [*sic*], as compared to around 2.1 million short falls among the same children over the same period. This means that the risk of death from long falls is correctly assessed by Chadwick as being around 1 in 118, whereas even if we assume that all the short fall histories were true, the short fall risk of death would be only about 7 in 2.1 million Chadwick's own data should be interpreted to tell us that if we accept the seven short fall histories as true, the risk of death from long falls is more than 2,500 times greater than the risk of death from short falls, which is totally opposite to his absurd assertion that we would be led to the conclusion that "the risk of death is eight times greater in children who fall from 1 to 4 feet than in those who fall from 10 to 45 feet."¹⁰⁵

Contrary to Dr. Narang's interpretation of the Chadwick study, one simply cannot conclude from the data that "short falls appear to be more dangerous than longer falls,"¹⁰⁶ or that it is therefore safe to disregard short-fall reports as implausible.

In a second frequently cited article on short falls, Chadwick and his colleagues attempted to calculate a rate of death for children from short falls based on the number of recorded short-fall deaths in the California Epidemiology and Prevention for Injury Control Branch (EPIC) database. To do so, they took the number of reported short-fall deaths among children and divided it by the total number of children in California, to arrive at a short-fall death rate of 0.48 deaths per million children¹⁰⁷—a figure used frequently to dismiss short-fall explanations in SBS/AHT prosecutions. Again, however, Chadwick's analysis suffers from serious statistical errors.

To start, Chadwick excluded almost half of the recorded short fall deaths (6 of 13) and did not fully explain why he excluded some of them.¹⁰⁸ Second, as Schneps has pointed out, Chadwick disregarded other empirical work showing much higher rates of short-fall deaths, without adequate justification.¹⁰⁹ These issues aside, however, both Chadwick's mode of calculation and the conclusions he draws from his calculation are demonstrably erroneous.

105. Schneps, *Paradoxes & Wrong Conclusions*, *supra* note 94, at 7.

106. Narang et al., *Part II*, *supra* note 98, at 219.

107. Chadwick et al., *Annual Risk of Death Resulting from Short Falls Among Young Children*, *supra* note 90, at 1213.

108. See Schneps, *Numerical & Reasoning Errors*, *supra* note 99, at 7.

109. See Schneps, *Paradoxes & Wrong Conclusions*, *supra* note 94, at 9-10.

There are at least three significant statistical errors in Chadwick's calculation. The first statistical error is Chadwick's reliance on what is likely incomplete and flawed data beyond the exclusions that Schneps noted. Among other problems with the dataset¹¹⁰ is the reality that it depends on the data submitters and recorders to properly determine that a child's death was attributable to a short fall in the first instance. Yet at the time that the data was collected, 1999–2003, the prevailing dogma espoused by the medical community was that short falls simply do not kill.¹¹¹ While the medical community now universally recognizes that this belief was wrong, it almost certainly meant that, when confronted with true short-fall deaths in that era, many physicians likely rejected that etiology and recorded the death as attributable to something else; the reported short-fall deaths in the 1999–2003 EPIC database almost surely represent a significant undercount of actual short-fall deaths.¹¹² As statistician Maria Cuellar observes, the data are subject to various forms of bias: "In some hospitals, some cases might be categorized as a short fall, and some cases might be categorized as shaken. That is a misclassification problem with the diagnosis."¹¹³

A second statistical problem identified by Cuellar is the implication drawn from Chadwick's work that a purportedly very low rate of short-fall deaths means physicians can disregard the possibility that a given child died from a short fall. For starters, as Cuellar observes, "rare events are not impossible."¹¹⁴ Beyond that, such a claim is not sustainable, statistically, unless data is also collected on the frequency or rarity of alternative explanations; standing alone, the data on short falls do not tell us what we need to know to assess the likelihood of a fall as an explanation in any given case, as op-

110. For a discussion of other problems with the dataset, see Maria Cuellar, *Short Fall Arguments in Court: A Probabilistic Analysis*, U. MICH. J. L. REFORM 763, 769 (2017).

111. For example, the official position paper of the American Academy of Pediatrics (AAP) in 2001 asserted unequivocally, "The constellation of these injuries [deemed diagnostic of SBS] does not occur with short falls[.]" Committee on Child Abuse and Neglect, American Academy of Pediatrics, *Shaken Baby Syndrome: Rotational Cranial Injuries—Technical Report*, 108 PEDIATRICS 206, 206 (2001). Given the incontrovertible evidence that has emerged since then, subsequent iterations of that position paper dropped the claim that short falls cannot kill. See Cindy W. Christian, Robert Block & the COMM. ON CHILD ABUSE & NEGLECT, AM. ACAD. OF PEDIATRICS, *Abusive Head Trauma in Infants and Children*, 123 PEDIATRICS 1409 (2009).

112. This methodological problem is a form of what is known as "detection bias." See Lyon & Koehler, *supra* note 64, at 68 (citing MICHAEL S. KRAMER, CLINICAL EPIDEMIOLOGY AND BIostatISTICS 53 (1988)).

113. Cuellar, *supra* note 110, at 771.

114. *Id.* at 765.

posed to some other explanation.¹¹⁵ Cuellar explains that this problem arises because Chadwick et al.'s short-fall death rate "is calculated in isolation . . . [T]hey calculate the probability that one event happens—that is, that a child will die from a short fall. But they do not calculate any of the probabilities for any other possible events that might have caused the outcomes, in order to provide a comparison of the two occurrences. We need to find out how likely these other possible causes are before we make any conclusions about what is more likely."¹¹⁶

Third, Chadwick's calculation uses the wrong denominator. Even if the data Chadwick utilized were correct—that only six children died from short falls in the study period—to put those six children in the numerator, and *all* children in the denominator, is grossly misleading. For when trying to determine cause of death in a particular case, we are already dealing with a rare occurrence—a child has died from a serious brain event or injury—and hence we are already dealing with a small subset of all children. Hence, the relevant inquiry is not what percentage of *all* children die from short falls, but what percentage of children who *have died from a brain injury* might have died from a short fall?¹¹⁷ Cuellar explains it this way:

[Chadwick and his colleagues] divided six by the number of all infants in California in that specific time period. But we know some more information about this infant, not just that he is a child in California. This is not a healthy child. This child has head trauma and has died. Or, in a different case, we might have a child who has not died, but also has head trauma. What we must do is restrict the population in light of this additional information. So, in a probability statement, we would write this as the probability that a child was shaken, given that this is an infant with head trauma and death. This "given" part is called "conditioning."¹¹⁸

115. *Id.* at 767.

116. *Id.*

117. If the question were a general one—what percentage of all children die from short falls?—then it would be appropriate to put the number of known short-fall deaths in the numerator, and *all* children in the denominator. But in a criminal case in which a child has died from brain injuries, a rare event has already occurred, and the important question becomes, of children who have suffered this kind of death, how many might be the result of a short fall?, then the denominator must include only those children who have died from brain injuries. That is the question that matters when assessing whether any particular death might have been produced by a short fall.

118. Cuellar, *supra* note 110, at 766.

If the number of known short-fall deaths (assuming that true figure is knowable) were divided not by *all* children, but rather by the relevant population—*all children who have died from brain injury*—the relevant rate of short fall deaths would be many times higher than the rate calculated by Chadwick. That is the figure that matters to physicians and courts tasked with determining what caused a child’s fatal brain injuries.

Moreover, even accepting Chadwick’s calculations, his data simply do not lead to the conclusion that the possibility of death from a short fall is so remote that it can be safely disregarded. Schneps has shown that, even using Chadwick’s estimated rate, with a population of about 200,000 children under age 5 in San Diego County in the period covered by his study, a 0.48% risk of a short fall death would translate into “about a 29% chance of seeing at least one legitimate short fall fatality in the period and population under study—a probability that is definitely not so small as to be negligible.”¹¹⁹ She concludes:

The error consists in assuming that it is incredibly unlikely to see an event that only occurs with a frequency of 0.48 per million children, forgetting that if one is actually observing a population containing a million children, it is actually quite likely that one will see one or two occurrences of the “rare” event.¹²⁰

And of course, if the true rate of short-fall deaths is higher than Chadwick estimates, as seems very likely, the chance of seeing some occurrences of this “rare” event is even higher—potentially significantly higher.¹²¹

These are just some of the errors in statistical thinking that can be found in the child abuse and forensic science literature. This essay is not meant to canvass all such errors. Others have examined additional serious problems with the statistical analyses in this arena.¹²² My purpose here is simply to highlight some of the recur-

119. Schneps, *Paradoxes & Wrong Conclusions*, *supra* note 94, at 11.

120. *Id.* at 12.

121. Schneps calculates, for example, that if one were to accept the short-fall-death rate calculated in another study of three short-fall deaths per million, the probability of seeing “exactly one fatality” in a population of 200,000 children under five (as in San Diego) would be about 25 percent, and the likelihood of seeing at least two fatalities over 3.5 years (Chadwick’s study period) is nearly 62 percent. *Id.* at 11-12 (citing A. Khambalia et al., *Risk Factors for Unintentional Injuries Due to Falls in Children Aged 0-6 Years: A Systematic Review*, 12 *INJ. PREVENTION* 378 (2006)).

122. See, e.g., Maria Cuellar, *Causal Reasoning and Data Analysis: Problems with the Abusive Head Trauma Diagnosis*, 16 *L., PROBABILITY & RISK* 223 (2017); Schneps, *Numerical & Reasoning Errors*, *supra* note 99.

ring problems, as illustrations, so as to alert physicians, forensic scientists, lawyers, and judges of the need to pay rigorous attention to proper application of statistics.

III. MAKING PROBABILISTIC CLAIMS THAT EXCEED THE AUTHORITY OF THE EXPERT AND INTRUDE ON THE PROVINCE OF THE JUDGE OR JURY

Statistics are not merely the domain of scientists and expert witnesses; they permeate legal thinking, even if lawyers, like forensic analysts, are not always aware of it. Indeed, the foundational principle in the Rules of Evidence—the rule of relevance in Federal Rule of Evidence 401¹²³—can be understood itself as an expression of Bayes’s Theorem. As discussed, Bayes’s Theorem teaches that the probability that an event or proposition is true (which of course is the legal fact-finder’s fundamental task) is determined by starting with prior odds (in the legal arena, whatever evidence has already been heard), multiplying those odds by the likelihood ratio presented by a new piece of evidence, and arriving at posterior odds—an updated assessment of likely guilt or innocence.¹²⁴

Again, the multiplier in this equation—the likelihood ratio—is simply the relationship of the probability of one hypothesis or conclusion divided by its opposite.¹²⁵ Applying this to forensic evidence in a criminal case, the numerator might be the odds that we would see a specific piece of evidence of interest found at the crime scene if the suspect was the source (here, this would be the prosecutor’s hypothesis), and the denominator would be the odds that we would see the evidence if someone other than the suspect was the source (here the defense hypothesis).¹²⁶ In this equation, logically,

123. FED. R. EVID. 401.

124. See *supra* notes 70–74 and accompanying text.

125. See *id.*

126. Likelihood ratios in pattern-matching disciplines are similar to those in SBS/AHT cases, with slight variation: they can be understood as the probability of seeing the evidence if the defendant is the source (the prosecutor’s hypothesis) divided by the probability of seeing the evidence if the defendant is not the source (the defense hypothesis). In statistical notation, the formula for the likelihood ratio is

$$\frac{\Pr(E|S)}{\Pr(E|\bar{S})}$$

where Pr means probability, E denotes the evidence, S means the suspect is the source of the evidence, and \bar{S} means suspect is not the source of the evidence. To put the likelihood ratio into Bayes’s Theorem, the formula is (prior odds X likelihood ratio = posterior odds):

$$\frac{\Pr(E|S)}{\Pr(E|\bar{S})}$$

any likelihood ratio of greater than one would support the proposition in the numerator—in this example, the prosecutor’s proposition that the evidence in question would be present if the defendant were the source. Any number less than one, that is, any fraction or decimal between zero and one, would support the proposition in the denominator—in this example, the defense proposition that the evidence would exist if it was left by someone other than the defendant. And a likelihood ratio of exactly one—that is, where the odds represented in the numerator and denominator are exactly the same—would not support either proposition over the other one.

Now consider the definition of relevant evidence under Rule 401:

Evidence is relevant if:

- (A) it has any tendency to make a fact more or less probable than it would be without the evidence; and
- (B) the fact is of consequence in determining the action.¹²⁷

Logically, therefore, evidence is relevant if its addition to the case produces a likelihood ratio about a fact of consequence that is any number other than one.¹²⁸ If the likelihood ratio produces a number that is any degree higher or lower than one, even the slightest bit, multiplying that number by the prior odds will change the posterior odds to some degree, thereby making the fact more or less probable than it appeared prior to application of this evidence. Understood this way, the statistical formulation helps us grasp the broad reach of the concept of relevancy under the Rules of Evidence.

The important point for this essay is to understand where in this formula the expert witness’s opinion fits. It is the role of the legal fact-finder (judge or jury) to reach ultimate conclusions about guilt or absence of guilt—what we might think of as an expression of the posterior odds. Theoretically, the legal fact-finder will embark on this task by considering the available relevant evidence in the case to intuit some belief in the prior odds of the defendant’s guilt. To that, the testimony of the expert (the forensic analyst) might add an evaluation of the evidence, i.e., a likelihood ratio—that is, an assessment of the likelihood of seeing the particular evidence (the fingerprints, the bullet striations, or the like) if the defendant is the source—which the fact-finder might use intuitively as

127. Fed. R. Evid. 401.

128. Richard Lempert’s classic 1977 article was perhaps the first to recognize this relationship between likelihood ratios and legal relevance under Rule 401. Lempert, *supra* note 70, at 1025.

a multiplier, to arrive at the posterior odds.¹²⁹ Breaking down the decision points required to assess the evidence in this way makes it clear that determining the prior odds and the posterior odds is, or least should be, a task reserved for the legal fact-finder, not the forensic analyst. The analyst only adds the likelihood ratio—the likelihood of seeing the particular forensic patterns if the defendant is the source.

It is in part for this reason that the NAS¹³⁰ and the short-lived National Commission on Forensic Science,¹³¹ and others,¹³² have warned against exposing forensic analysts to context information unrelated to the forensic analysis that might bias the analysis one way or another (thereby leading the analyst inappropriately and probably subconsciously to start with an assessment of prior odds).¹³³ The analyst's proper role, of providing just the likelihood or relevance ratio derived from the forensic analysis, is also why forensic analysts should not be permitted to opine about posterior odds on the question of guilt—that is, the ultimate question about how likely it is the defendant is the perpetrator.¹³⁴ Hence, forensic

129. . There is a debate in the literature about whether ordinary people actually think like Bayesians. For a description of this debate and citations to the relevant literature, see William C. Thompson, *How Should Forensic Scientists Present Source Conclusions?*, 48 Seton Hall L. Rev. 773, 804–05 (2018). Regardless of how lay fact-finders actually weigh evidence, however, the Bayesian model does work as a logical descriptor of the theory underlying the Rules of Evidence.

130. NAS Report, *supra* note 4, at 24.

131. Nat'l Comm'n on Forensic Sci., *Views of the Commission: Ensuring That Forensic Analysis is Based Upon Task-Relevant Information* (2015).

132. See, e.g., William C. Thompson, *Determining the Proper Evidentiary Basis for an Expert Opinion: What Do Experts Need to Know and When Do They Know Too Much?*, in *Blinding as a Solution to Bias: Strengthening Biomedical Science, Forensic Science, and Law* 133 (Christopher T. Robertson & Aaron S. Kesselheim eds., 2015); Saul M. Kassin, Itiel E. Dror & Jeffrey Kukucka, *The Forensic Confirmation Bias: Problems, Perspectives, and Proposed Solutions*, 2 J. App. Research in Memory & Cognition 42 (2013); D. Michael Risinger et al., *The Daubert/Kumho Implications of Observer Effects in Forensic Science: Hidden Problems of Expectation and Suggestion*, 90 Cal. L. Rev. 1 (2002); Keith A. Findley & Michael Scott, *The Multiple Dimensions of Tunnel Vision in Criminal Cases*, 2006 Wis. L. Rev. 291 (2006).

133. Underlying these recommendations to separate forensic analysts from “task-irrelevant” contextual information is that such “task-irrelevant” contextual information will change the way the experts interpret the physical evidence, and hence distort assessment of its value. In other words, the concern is that contextual bias will distort the likelihood ratio. Consideration of task-irrelevant context information thus runs two risks: it can contaminate the analyst's assessment of the likelihood ratio, and it exceeds the expert's authority and intrudes on the province of the jury to assess prior odds and posterior odds.

134. While Federal Rule of Evidence 704(a) generally provides that “[a]n opinion is not objectionable just because it embraces an ultimate issue,” the ultimate issue of guilt or innocence should never be permissible because that would

analysts are typically permitted to testify about whether evidentiary samples “are consistent” or even are a “match” (as problematic as that is as well), or even more problematically, that the defendant is the “source” of crime scene evidence or that a child was abused, but they should not be permitted to opine that therefore the accused is guilty.¹³⁵

While that might all seem obvious enough, this concept gets more challenging when examined more deeply. Consider, for example, the ways that context evidence can sneak into a forensic analyst’s conclusions. As noted, when an analyst is exposed to task-irrelevant context evidence (e.g., a fingerprint analyst is told that the suspect was identified by a witness, or confessed), that information can bias the way the analyst interprets the fingerprint evidence.¹³⁶ Beyond that, some analysts report their findings as “source probabilities,” that is, “conclusions about the probability that the items have a common source, which can be expressed, either with numbers (e.g., ‘there is a 99% chance this bite mark was made by the suspect’) or with words (‘it is highly probable that these marks were made by the same tool’).”¹³⁷ But as William Thompson has observed, opinions about “source probabilities” are problematic because they are “based partly on the examiner’s analysis of the physical characteristics of the items being compared, and partly on the examiner’s assumptions or conclusions about the strength of other evidence that bears on whether the items have a common source”¹³⁸—that is, on the task-irrelevant information that is reserved for the fact-finders in assessing the prior odds and posterior odds of guilt.

This problem can be even more pronounced with other types of forensic expert evidence. Consider, again, medical opinions

invade the province of the jury and would require the expert to consider all of the evidence in the case, beyond merely the information that went into the forensic analysis, and would cover all of the elements of the offense.

135. See, e.g., *United States v. Wright*, 48 M.J. 896, 901–02 (A.F. Ct. Crim. App. 1998) (“Expert testimony may not be used to determine the credibility of the victim nor may an expert offer an opinion as to the guilt or innocence of the accused.”); *Stephens v. State*, 774 P.2d 60, 66 (Wy. 1989) (quoting 3 Charles E. Torcia, *Wharton’s Criminal Evidence* § 566 (14th ed. 1987) (“[A] witness may not state his opinion as to . . . whether the defendant was guilty or innocent of the crime charged[.]”)); *United States v. Thanh Quoc Hoang*, 891 F. Supp. 2d 1355, 1362 (M.D. Ga. 2012) (quoting *Montgomery v. Aetna Cas. & Sur. Co.*, 898 F.2d 1537, 1541 (11th Cir. 1990) (“Although Rule 704(a) abolished the ultimate issue rule, an expert ‘may not, however, merely tell the jury what result to reach. A witness also may not testify to the legal implications of conduct.’”)).

136. See Risinger et al., *supra* note 132.

137. Thompson, *supra* note 129, at 777.

138. *Id.* at 809.

about cause and manner of injury or death, and in particular the example of medical opinions about child abuse. As we have seen, when determining the significance of diagnostic findings, it is important for physicians to take into account the likelihood ratio and the base rate (the prior odds) of the particular condition or event in question. Hence, as discussed, when assessing the significance of subdural hematoma with regard to abuse, the physician must consider the base rate of abuse in a population.¹³⁹ Likewise, when considering the *rate* of short-fall deaths in a population, the physician must consider the base rate of short falls in a population.

It is quite another matter, however, for the physician to consider the prior odds of *guilt*, or to opine about the posterior odds of *guilt*. That is what our system reserves for the judge or jury as fact-finder. On the ultimate question of guilt or innocence, the expert physician is still appropriately limited to providing only the likelihood ratio—the likelihood of seeing the particular medical findings if they were produced by abuse, over the likelihood of seeing the medical findings if they were produced by natural or accidental causes, based upon scientific measurements. Understood this way, it becomes clear why factors that a judge or jury is capable of considering without expert assistance, and which the judge or jury will likely consider that relate to the ultimate issue of guilt or innocence (the prior and posterior odds), should not be included as considerations in the physician's opinion, at least when those opinions are offered in the courtroom where the judge or jury is the ultimate fact-finder.¹⁴⁰ Yet it is not unusual for child abuse physicians to consider, in their subjective, clinical-judgment-based assessments of whether a child was abused, such matters as the demeanor of the caregivers, the believability of their denials of guilt, the quality of their home, the nature of the parental relationship, whether the

139. See *supra* notes 74–85 and accompanying text.

140. In this regard it is important to note that the context in which an expert is asked to render an opinion can affect what evidence the expert might appropriately consider. When, for example, a medical examiner is required by statute to fill out a death certificate that includes a determination of both cause and manner of death, the medical examiner must consider all relevant evidence—medical and non-medical—to determine, for example, if a death was a homicide, suicide, accident, natural event, or undetermined. But in that context, by law, the medical examiner is acting as the ultimate fact-finder. See NAS Report, *supra* note 4, at 243–44. In the courtroom, however, the jury (or judge) is the ultimate fact-finder. In that setting, the medical examiner's job, as an expert under the Rules of Evidence, see Fed. R. Evid. 702, is to apply particular expertise that the jury is otherwise incapable of applying itself, but it should not be to render opinions that depend on all of the other evidence in the case that the jury *is* capable of assessing itself.

caregiver confessed, or even the race of the parties.¹⁴¹ But properly understood, those factors are *other* evidence in the case, the kinds of evidence that inform the judge or jury's intuitive assessment of prior odds or posterior odds. Therefore, if the the physician were also to rely on such other evidence to assess prior and posterior odds, it would both invade the province of the jury and lead to subtle double-counting of those factors (first by the expert, then by the jury).

Yet, based on assessments of both prior odds and likelihood ratios, it is not unusual for the child-abuse physician to opine as well on posterior odds, concluding that a child was indeed shaken or shaken and slammed and that the force applied had to have been so immense it could not have been accidental. In this way, the expert fully occupies the field, opining about both the *actus reus* and the *mens rea* for the crime of child abuse, essentially providing what Deborah Tuerkheimer has aptly called a “medical diagnosis of murder.”¹⁴² Virtually no other expert is permitted to so fully assess guilt or innocence for the jury (or judge). Breaking down the component parts of the factors that go into this assessment should make it clear that such broad claims require probabilistic judgments of non-forensic considerations that go well beyond what an expert in our system should be permitted to assess.

CONCLUSION

William Thompson, Joëlle Vuille, Franco Taroni, and Alex Biedermann predicted recently, “[w]e are thus on the cusp of a new era for forensic science—an era in which statistics will inevitably play a greater role.”¹⁴³ That prediction is certainly correct; indeed, that future is here. And for that we should be glad, because with more attention to proper application and presentation of statistics, some of the deficiencies in forensic analysis and expert testimony

141. Recent research by Itiel Dror and his colleagues has highlighted some of the serious risks of biased interpretation that can be created when, for example, forensic pathologists are exposed to context information that has no relevance to their *medical* analysis of cause and manner of death in child death cases. Dror et al. found that, when pathologists were presented with information suggesting that a deceased child was Black and died in the care of the the mother's boyfriend, the pathologists were many more times more likely to conclude that the case involved a homicide as opposed to an accident, than when they were told that the child was white and died while in the care of the child's grandmother. Itiel Dror, Judy Melinek, Jonathan L. Arden, Jeff Kukucka, Sarah Hawkins, Joye Carter & Daniel S. Atherton, *Cognitive Bias in Forensic Pathology Decisions*, J. Forensic Sci. 1 (2021).

142. Deborah Tuerkheimer, *The Next Innocence Project: Shaken Baby Syndrome and the Criminal Courts*, 87 Wash. U. L. Rev. 1, 5 (2009).

143. Thompson et al., *supra* note 18, at 27.

might be exposed and then corrected or prevented. Claims presented in categorical ways can be exposed as probabilistic and can be reframed accordingly. Probabilistic claims without supporting data can be exposed and their limitations explained, or the claims excluded. Statistical errors can be corrected, so data is properly interpreted and presented to fact-finders. And the separate roles of expert witness and jury (or judge) can be more clearly understood, and the evidence relevant to the task of each can be appropriately apportioned. But all of this will depend on the vigilance of experts, lawyers, and judges¹⁴⁴ alike to properly apply probabilities and to ensure that the statistics inform, rather than obfuscate.

144. Jonathan J. Koehler, *How Trial Judges Should Think About Forensic Science Evidence*, 102 JUDICATURE 28, 36 (2018) (arguing that it is ultimately up to the courts to make forensic science more scientific because experts and lawyers will not make the necessary changes unless the changes are forced on them from the outside).